Visual Reasoning Strategies for Effect Size Judgments and Decisions

Alex Kale, Matthew Kay, and Jessica Hullman

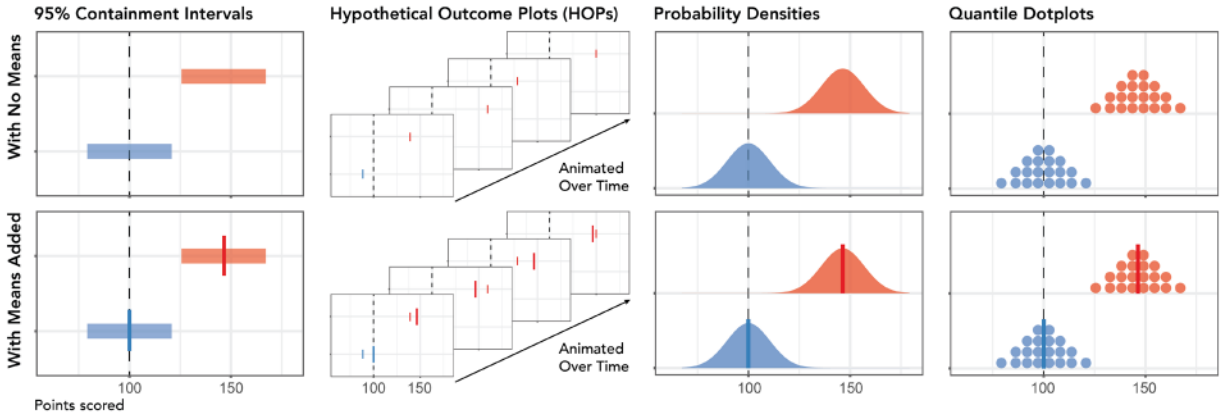


Fig. 1: Visualization designs evaluated in our experiment.

Abstract— Uncertainty visualizations often emphasize point estimates to support magnitude estimates or decisions through visual comparison. However, when design choices emphasize means, users may overlook uncertainty information and misinterpret visual distance as a proxy for effect size. We present findings from a mixed design experiment on Mechanical Turk which tests eight uncertainty visualization designs: 95% containment intervals, hypothetical outcome plots, densities, and quantile dotplots, each with and without means added. We find that adding means to uncertainty visualizations has small biasing effects on both magnitude estimation and decision-making, consistent with discounting uncertainty. We also see that visualization designs that support the least biased effect size estimation do not support the best decision-making, suggesting that a chart user's sense of effect size may not necessarily be identical when they use the same information for different tasks. In a qualitative analysis of users' strategy descriptions, we find that many users switch strategies and do not employ an optimal strategy when one exists. Uncertainty visualizations which are optimally designed in theory may not be the most effective in practice because of the ways that users satisfice with heuristics, suggesting opportunities to better understand visualization effectiveness by modeling sets of potential strategies.

Index Terms—Uncertainty visualization, graphical perception, data cognition

1 Introduction

Many visualization authors perceive visualizing uncertainty as an exception, rather than a norm [25]. However, the common practice of omitting uncertainty information from visualizations and focusing attention on point estimates leads to "incredible certitude" [38, 39], the unwarranted impression that error is minimal or not important. To enable informed judgments and decisions, a common suggestion is to present uncertainty information alongside point estimates, for example, by showing intervals in which estimates could fall [11, 12, 37, 52].

However, presenting uncertainty alongside point estimates may not lead users to incorporate uncertainty information into their judgments. A large body of work on biases due to heuristics (e.g., [30, 54, 55]), also commonly known as *satisficing* [45], shows that people often avoid or discount uncertainty information. This suggests that chart users may ignore uncertainty in favor of means even when both are presented [26].

Different visualization design choices make the mean more or less salient. Imagine a continuum of uncertainty visualization designs representing how perceptually difficult it is to decode the mean from a chart. At one extreme are hypothetical outcome plots or HOPs [26, 33] where the mean is only encoded implicitly as the average of a set of outcomes presented across frames of an animation. At the other extreme are direct encodings of point estimates presented alongside uncertainty (e.g., represented as error bars). We expect that the salience of the mean in uncertainty visualization designs and other factors such as frequency-framing of probability [14, 26, 33, 34] influence the degree to which users focus on means and ignore uncertainty.

How might chart users who focus on means judge effect size? Imagine a user viewing visualizations like those in Figure 1. Discounting uncertainty may manifest as using *distance between means or gist estimates of distance between distributions as a proxy for effect size* and not judging distance relative to the width of distributions. Using only distance as a proxy for effect size may be misleading (Fig. 2) because the distance between distributions depends on a number of factors, including the variance of distributions and the visualization author's choice of axis scale as noted by previous work [9, 21, 61].

We investigate a scenario where distance heuristics lead to a predictable pattern of bias in order to measure how different visualization designs impact users' reliance on distance as a proxy for effect size. Users are shown charts depicting various effects on a fixed axis (Fig. 2) such that when distributions have lower variance, visual distance between means is small regardless of effect size, but distances correspond to effect size more consistently at higher variance. In this scenario, we expect that adding means to uncertainty visualizations leads users to

- *Alex Kale is with the University of Washington. E-mail: kalea@uw.edu.*
- *Matthew Kay is with the University of Michigan. E-mail: mjskay@umich.edu.*
- *Jessica Hullman is with Northwestern University. E-mail: jhullman@northwestern.edu.*

Manuscript received 30 Apr. 2020; revised 31 July 2020; accepted 14 Aug. 2020. Date of publication 13 Oct. 2020; date of current version 15 Jan. 2021. Digital Object Identifier no. 10.1109/TVCG.2020.3030335

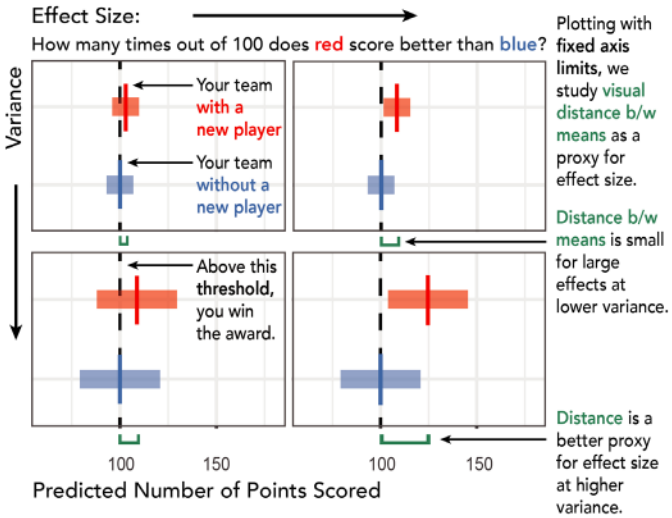


Fig. 2: Intervals with means showing two levels of effect size (72% and 95% Pr(S)) at low and high variance. Using visual distance between means as a proxy for effect size should result in greater bias toward underestimating effect size at lower variance than at higher variance.

underestimate effect size at lower variance. Conversely, adding means may reduce this underestimation bias at higher variance.

We contribute a pre-registered experiment on Mechanical Turk investigating how uncertainty visualization design impacts lay users' judgments and decisions from effect size. We find that *visualization designs which support magnitude estimation are not necessarily best suited as decision aids*. Quantile dotplots lead to the least bias in magnitude estimation, but other visualizations lead to the least bias in decision-making. On a fixed axis scale, densities without means support unbiased decisions at lower variance, and users show substantial bias with all visualizations at higher variance. Visualization effectiveness for decision-making depends on the level of variance in data relative to the axis scale. Adding means has a negligible impact on magnitude estimation, but in most cases it leads to less utility-optimal decisions.

In a qualitative analysis of users' strategy descriptions, we find that *few users apply the optimal strategy* for reading an uncertainty visualization when one exists. Instead, the majority of users appear to *satisfice* [45] by using a small set of heuristics. We find that the *majority of users report relying on visual distance* between distributions regardless of uncertainty information, an observation that is consistent with the biases in our quantitative results. We also find that many users switch between strategies. This suggests that many uncertainty visualizations may not be interpreted in ways that researchers and designers expect, and characterizing possible strategies may lead to design recommendations based on how users reason in practice.

2 Background: Visualizing Uncertainty

In communicating the results of statistical analysis, visualization authors commonly represent uncertainty as a range of possible values as recommended by numerous experts (e.g., [11, 37, 52]). Other conventional uncertainty representations commonly used in statistical analysis include aggregate encodings of distributions such as boxplots [53], histograms [42], and densities [2, 48]. Frequency-based uncertainty visualizations build on a large body of work suggesting that framing probabilities as frequencies of events improves statistical reasoning [6, 14, 16, 17, 20, 22, 26, 33, 34, 36, 41]. These include hypothetical outcome plots (HOPs) [26], which encode possible outcomes as frames in an animation, and quantile dotplots [34], which quantize a distribution of possible outcomes and represent each quantile as a discrete dot. A growing body of work suggests that lay and expert audiences commonly misinterpret interval representations of uncertainty [3, 19, 47] and that other uncertainty visualization formats such as gradient plots [10], violin plots [10, 26], HOPs [26, 33], and quantile dotplots [14, 34] lead to more accurate interpretation and performance on various tasks.

In our study, we compare two frequency-based visualizations, quantile dotplots and HOPs, with two more conventional uncertainty representations, intervals and densities. By testing each with and without added means, we investigate the extent to which users of these uncertainty visualizations differ in their tendency to ignore uncertainty.

When chart users don't know how to interpret uncertainty, prior work [26] suggests that they may substitute a judgment of the mean difference between distributions for more complicated judgments about the reliability of effects. This *visual distance heuristic* motivates design principles, for example, that the quantitative axis on a bar chart should always start at zero [4, 24], or that axis scales should align visual distance with effect size [61]. Axis scale impacts the perceived importance of effect size regardless of chart type (e.g., lines versus bars) and despite attempts to signal that an axis does not start at zero (e.g., breaking the axis) [9]. Rescaling the axis on a chart that displays inferential uncertainty (e.g., 95% confidence intervals) to the scale implied by descriptive uncertainty (e.g., 95% predictive intervals) can reduce bias in impressions of effect size [21]. In our study, we investigate the visual distance heuristic by asking users to compare distributions with different levels of variance on a common scale (Fig. 2).

3 Method

We tested how adding means to different uncertainty visualizations impacts users' estimates and incentivized decisions from effect size.

3.1 Tasks & Procedure

Our task was like a fantasy sports game. We showed participants charts comparing the predicted number of points scored by their team with and without adding a new player (e.g., Fig. 2). Participants estimated the effect size of adding the new player and decided whether or not to pay to add the new player to their team.

Effect Size Estimation: We asked participants to estimate a measure of effect size called *probability of superiority* or common language effect size [40]: *"How many times out of 100 do you estimate that your team would score more points with the new player than without the new player?"* We elicited probabilities as "times out of 100" based on literature in statistical reasoning (e.g., [17, 20]) suggesting that people reason more accurately with probabilities when they are framed as frequencies. Probability of superiority, the percent of the time that outcomes for one group A exceed outcomes for another group B, is a proxy for standardized mean difference $\frac{\mu_A - \mu_B}{\sigma_{A-B}}$ [8, 13], the difference between two group means relative to uncertainty in the estimates. Using synthetic data (see Section 3.5), we evaluated bias in effect size estimates compared to a known ground truth.

Intervention Decisions: We also asked participants to make binary decisions indicating whether they would *"Pay for the new player,"* or *"Keep [their] team without the new player."* On each trial, the participant's goal was to win an award worth $3.17M, and they could pay $1M to add a player to their team if they thought the new player improved their chances of winning enough to be worth the cost. There were four possible payouts in each trial:

1. The participant won without paying for a new player (+$3.17M).
2. The participant paid for a new player and won (+$2.17M).
3. They failed to win without paying for a new player ($0).
4. The participant paid for a new player and failed to win (-$1M).

The user could only lose money if they paid for the new player.[1] We set up the incentives for our task so that a risk-neutral chart user should pay for a new player only when effect size was larger than 74% probability of superiority or Cohen's d of 0.9, the average effect size in a recent survey of studies in experimental psychology [44]. This enabled us to evaluate intervention decisions compared to a utility-optimal standard.

Feedback: At the end of each trial we told users whether or not their team scored enough points to win an award, using a Monte Carlo simulation to generate a win or loss based on the participant's decision.

[1] In pilot studies, we tested how framing outcomes as winning versus losing awards impacted user behavior and found that participants had greater preference for intervention when it was described as increasing the certainty of gains, consistent with prior work by Tversky and Kahneman [31, 55].

We split feedback into two tables. One showed the change in account value for the current trial. The other showed cumulative account value and how this translated into a bonus in real money. By showing probabilistic outcomes, instead of the expected value of decisions, feedback gave participants a noisy signal of how well they were doing, mirroring real-world learning conditions for decisions under uncertainty.

Payment: Participants received a *guaranteed reward of $1* plus a bonus of $\$0.08 \cdot (account - \$150M)$, where \$0.08 per \$1M was the exchange rate from account value to real dollars, *account* was the value of their fantasy sports account at the end of the experiment, and \$150M was a cutoff account value below which they receive no bonus. These values were carefully chosen to result in bonuses between \$0 and \$3, such that participants who guessed randomly and experienced unlucky probabilistic outcomes would receive no bonus, and participants who responded optimally would be guaranteed a bonus.

User Strategies: To supplement our quantitative measures with qualitative descriptions of users' visual reasoning, at the end of each of the two block of trials, we asked users, *"How did you use the charts to complete the task? Please do your best to describe what sorts of visual properties you looked for and how you used them."*

3.2 Formalizing a Class of Decision Problems

Our decision task represents a class of decision problems where one makes a *binary decision* about whether or not to invest in an intervention that changes the probability of an *all-or-nothing outcome*. For example, this class of problems includes medical decisions about treatments that may save someone's life or cure them of a disease, organizational decisions about hiring personnel to reach a contract deadline, and personal decisions such as paying for education to seek a promotion. Previous decision-making literature examines similar problems in the context of salting the road in freezing weather [28, 29], voting in presidential elections [59], and willingness to pay for interventions in a fictional scenario [21]. The key similarity between these decision problems is that their incentive structures imply a *common utility function*.

A utility function defines optimal (i.e., utility maximizing [57]) decisions for a risk-neutral observer, providing a normative benchmark used to measure bias in decision-making. Comparing behavior to a risk-neutral benchmark is a common practice in judgment and decision-making studies [1], often used to measure risk preferences [58] or attitudes that make a person more or less inclined to take action than they should be based on a cost-benefit analysis. In the class of decision problems we investigate, the implied utility function depends on both the amount of money one stands to win or lose (e.g., the value of an award and the cost of a new player) and the effect size (e.g., the difference in team performance with versus without a new player).

Let v be the value of an award. Let c be the cost of adding a new player to the team. The utility-optimal decision rule is to intervene if

$$v \cdot \Pr(award|\neg player) < v \cdot \Pr(award|player) - c$$

where $\Pr(award|\neg player)$ is the probability of winning an award *without* a new player, and $\Pr(award|player)$ is the probability of winning an award *with* a new player. Assuming a constant ratio between the value of the award and the cost of intervention $k = \frac{v}{c}$, we express the decision rule in terms of the difference between the probabilities of winning an award with versus without a new player:

$$\Pr(award|\neg player) + \frac{1}{k} < \Pr(award|player)$$

The threshold level of effect size above which one should intervene depends on the incentive ratio k and the probability of a payout without intervention $\Pr(award|\neg player)$. In our study, we fixed the incentives $k = 3.17$ and the probability of winning an award without a new player $\Pr(award|\neg player) = 0.5$ so that users would not have to keep track of changing incentives, and effect size alone was the signal that users should base decisions on.[2] This enabled a controlled evaluation of how users translate visualized effect size into a sense of utility. By modeling a functional relationship between effect size and utility, we go beyond prior work which either does not vary the effectiveness of interventions (e.g., [28, 29, 59]) or examines only two levels of effect size as a robustness check for statistical tests (e.g., [21]).

3.3 Experimental Design

We assigned each user to one of four uncertainty visualization conditions at random, making comparisons of uncertainty visualizations between-subjects. On each trial, users made a probability of superiority estimate and an intervention decision. We asked users to make repeated judgments for two blocks of 16 trials each. In one block, we showed the users visualizations with means added, and in the other block there were no means. We counterbalanced the order of these blocks across participants. Each of the 16 trials in a block showed a unique combination of ground truth effect size (8 levels) and variance of distributions (2 levels), making our manipulations of ground truth, variance, and adding means all within-subjects. The order of trials in each block was randomized. In the middle of each block, we inserted an attention check trial, later used to filter participants who did not attend to the task. Users always saw an attention check at 50% probability of superiority with means and at 99.9% without means. Hence, each participant completed 17 trials per block and 34 trials total.

3.4 Uncertainty Visualization Conditions

We evaluated visualizations intended to span a design space characterized by the visual salience of the mean, expressiveness of uncertainty representation, and discrete versus continuous encodings of probability. As described above, we showed four uncertainty visualization formats—intervals, hypothetical outcome plots (HOPs), density plots, and quantile dotplots—with and without separate (i.e., extrinsic) vertical lines encoding the mean of each distribution. We expected that adding means would bias effect size estimates toward discounting uncertainty and that this effect would be most pronounced for uncertainty visualizations in which the mean is *not* intrinsically salient.

Intervals: We showed users intervals representing a range containing 95% of the possible outcomes (Fig. 1, left column). In the absence of a separate mark for the mean, the mean was not intrinsically encoded, and the user could only find the mean by estimating the midpoint of the interval. Intervals were not very expressive of probability density since they only encoded lower and upper bounds on a distribution.

Hypothetical Outcome Plots (HOPs): We showed users animated sequences of strips representing 20 quantiles sampled from a distribution of possible outcomes (Fig. 1, left center column), matching the data shown in quantile dotplots. Animations were rendered at *2.5 frames per second with no animated transitions* (i.e., tweening or fading) between frames, looping every 8 seconds. We shuffled the two distributions of 20 quantiles using a 2-dimensional quasi-random Sobol sequence [46] to minimize the apparent correlation between distributions. Like intervals, HOPs did not make the mean intrinsically salient, as means were implicitly encoded as the average position of an ensemble of strips shown over time. However, HOPs were more expressive of the underlying distribution than intervals and expressed uncertainty as frequencies of events, so they conveyed an experience-based sense of probability.

Densities: We showed users continuous probability densities where the height of the area marking encoded the probabilities of corresponding possible outcomes on the x-axis (Fig. 1, right center column). Unlike intervals and HOPs, the mean was explicitly represented as the point of maximum mark height because distributions were symmetrical, so means were intrinsically salient. Densities were also the most expressive of the underlying probability density function among the uncertainty visualizations we tested.

Quantile Dotplots: We showed users dotplots where each of 20 dots represented a 5% chance of a corresponding possible outcome on the x-axis (Fig. 1, right column). Like densities, because distributions were symmetrical and dots were stacked in bins to express this symmetry,

[2] In pilot studies, we tried manipulating k and $\Pr(award|\neg player)$ and found that these changes had little impact on the effectiveness of different uncertainty visualizations for supporting utility-optimal decision-making. In light of prior work showing that Mechanical Turk workers do not respond to changes of incentives [50], we suspect that these manipulations might have an impact in real-world settings which is difficult to measure on crowdsourcing platforms.

the mean was explicitly represented as the point of maximum height and was thus intrinsically salient.

3.5 Generating Stimuli

We generated synthetic data covering a range of effect size, so there were an equal number of trials where users should and should not intervene. Recall that 50% corresponded to a new player who did not improve the team's performance at all, 100% corresponded to a definite improvement in performance, and 74% was the utility-optimal decision threshold. We sampled eight distinct levels of ground truth probability of superiority, four values between 55% and 74% and four values between 74% and 95%, such that there are an equal number of trials above and below the utility-optimal decision threshold. Prior work in perceptual psychology [18, 62] suggests that the brain represents probability on a log odds scale. For this reason, we converted probabilities into log odds units and sampled on this logit-transformed scale using linear interpolation between the endpoints of the two ranges described above. We added two attention checks at probabilities of superiority of 50% and 99.9%, where the decision task should have been very easy, to allow for excluding participants who were not paying attention.

To derive the visualized distributions from ground truth effect size, we made a set of assumptions. We assumed *equal and independent variances* for the distributions with and without a new player σ^2_{team} such that $\sigma^2_{diff} = 2\sigma^2_{team}$ where σ^2_{diff} was the variance of the difference between distributions. We tested two levels of variance, setting the standard deviation of the difference between distributions σ_{diff} to a low value of 5 or a high value of 15. These levels produced distributions that looked relatively narrow or wide compared to the width of the chart, making visual distance between distributions an unreliable cue for effect size such that at low variance large effect sizes corresponded to distributions that looked close together.

We determined the distance between distributions, or mean difference μ_{diff}, using the formula $\mu_{diff} = d \cdot \sigma_{diff}$ where d were ground truth values as standardized mean differences (i.e., Cohen's d [8, 13]). The mean number of points scored without the new player was held constant $\mu_{without} = 100$, which corresponded to a 50% chance of winning the award. We calculated mean for the team with a new player $\mu_{with} = \mu_{without} + \mu_{diff}$. We rendered our chart stimuli using the parameters μ_{with}, $\mu_{without}$, and σ_{team} to define the two distributions on each chart. Holding the chance of winning without a new player constant at 50% (Fig. 2, blue distributions) is an experimental control that enables us to compare a user's preference for new players across trials using a coin flip gamble as the alternative choice, which is common in judgment and decision-making studies [1].

3.6 Modeling

We wanted to measure how much users underestimate effect size in their probability of superiority responses, how much they deviate from a utility-optimal criterion in their decisions, and how sensitive they are to effect size for the purpose of decision-making. To measure underestimation bias, we fit a linear in log odds model [18, 62] to probability of superiority responses, and we derive *slopes* describing users' responses as a function of the ground truth (Fig. 3). To measure bias and sensitivity to effect size in decision-making, we fit a logistic regression to intervention decisions, and we derive *points of subjective equality* and *just-noticeable differences* describing the location and scale of the logistic curve as functions of effect size (Fig. 4).

3.6.1 Approach

We used the brms package [5] in R to build Bayesian hierarchical models for each response variable: probability of superiority estimates and decisions of whether or not to intervene. We started with simple models and gradually added predictors, checking the predictions of each model against the empirical distribution of the data. This process of *model expansion* [15] enabled us to understand the more complex models in terms of how they differ from simpler ones.

We started with a *minimal model*, which had the minimum set of predictors required to answer our research questions, and built toward a *maximal model*, which included all the variables we manipulated in our experiment. We specified the minimal and maximal models for each response variable in our preregistration.[3]

Expanding models gradually helped us determine priors one-at-a-time. Each time we added a new kind of predictor to the model (e.g., a random intercept per participant), we honed in on weakly informative priors using prior predictive checks [15]. We centered the prior for each parameter on a value that reflected no bias in responses. We scaled each prior to avoid predicting impossible responses and to impose enough regularization to avoid issues with convergence in model fitting. We documented priors and model expansion in Supplemental Materials.[4]

3.6.2 Linear in Log Odds Model

We use the following model (Wilkinson-Pinheiro-Bates notation [5, 43, 60]) for responses in the probability of superiority estimation task:

$$\begin{aligned}
\text{logit}(response_{\Pr(S)}) \sim& \text{Normal}(\mu, \sigma)\\
\mu =& \text{logit}(true_{\Pr(S)}) * means * var * vis * order\\
&+ \text{logit}(true_{\Pr(S)}) * vis * trial\\
&+ \big(\text{logit}(true_{\Pr(S)}) * trial + means * var \,\big|\, worker\big)\\
\log(\sigma) =& \text{logit}(true_{\Pr(S)}) * vis * trial\\
&+ means * order\\
&+ \big(\text{logit}(true_{\Pr(S)}) + trial \,\big|\, worker\big)
\end{aligned}$$

Where $response_{\Pr(S)}$ is the user's probability of superiority response, $true_{\Pr(S)}$ is the ground truth probability of superiority, *trial* is an index of trial order, *means* is an indicator for whether or not extrinsic means are present, *var* is an indicator for low versus high variance, *vis* is a dummy variable for uncertainty visualization condition, *order* is an indicator for block order, and *worker* is a unique identifier for each participant used to model random effects. Note that there are submodels for the mean μ and standard deviation σ of user responses.

Motivation: We apply a logit-transformation to both $response_{\Pr(S)}$ and $true_{\Pr(S)}$, changing units from probabilities of superiority into log odds, because prior work suggests that the perception of probability should be modeled as linear in log odds (LLO) [18, 62]. We model effects on both μ and σ because we noticed in pilot studies that the spread of the empirical distribution of responses varies as a function of the ground truth, visualization design, and trial order. However, we are most interested in effects on mean response. The term $\text{logit}(true_{\Pr(S)}) * means * var * vis * order$ tells our model that the slope of the LLO model varies as a joint function of whether or not means were added, the level of variance, uncertainty visualization, and block order (i.e., all of these factors interacted with each other). This enables us to answer our core research questions, while controlling for order effects. The term $\text{logit}(true_{\Pr(S)}) * vis * trial$ models learning effects, so we isolate the impact of uncertainty visualizations. In both submodels, we added within-subjects manipulations as random effects predictors as much as possible without compromising model convergence.

3.6.3 Logistic Regression

We use this model to make inferences about intervention decisions:

$$\begin{aligned}
intervene \sim& \text{Bernoulli}(p)\\
\text{logit}(p) =& evidence * means * var * vis * order\\
&+ evidence * vis * trial\\
&+ \big(evidence * means * var + evidence * trial \,\big|\, worker\big)
\end{aligned}$$

Where *intervene* is the user's choice of whether or not to intervene, p is the probability that they intervene, and *evidence* is a logit-transformation of the utility-optimal decision rule (see Section 3.2):

$$evidence = \text{logit}(\Pr(award|player)) - \text{logit}(\Pr(award|\neg player) + \frac{1}{k})$$

This gives us a uniformly sampled scale of evidence where zero represents the utility-optimal decision threshold. All other factors are the same as in the LLO model (see Section 3.6.2).

[3] https://osf.io/9kpmb

[4] https://github.com/kalealex/effect-size-jdm

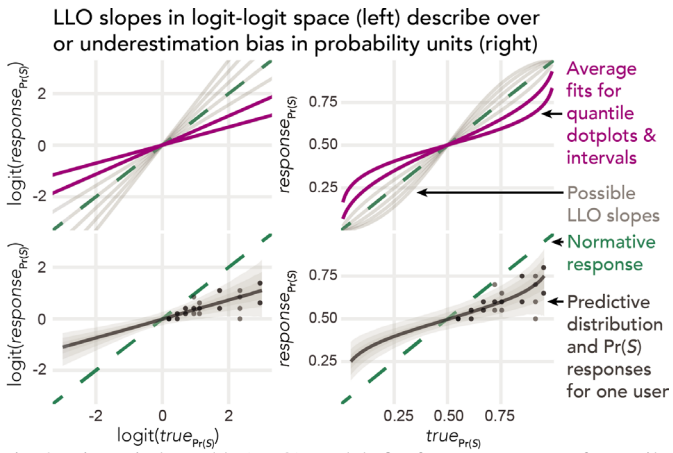


Fig. 3: Linear in log odds (LLO) model: fits for average user of quantile dotplots and intervals compared to a range of possible slopes (top); predictive distribution and observed responses for one user (bottom).

Motivation: We logit-transform our evidence scale because internal representations of probabilities are thought to be on a log odds scale [18, 62], such that linear changes in log odds appear similar in magnitude. The term $evidence * means * var * vis * order$ tells our model that the location and scale of the logistic curve vary as a joint function of whether or not means were added, the level of variance, uncertainty visualization, and block order. Mirroring an analogous term in the LLO model, this enables us to answer our core research questions, while controlling for order effects. The term $evidence * vis * trial$ models learning effects. As with the LLO model, we specify random effects per participant through model expansion by trying to incorporate as many within-subjects manipulations as possible.

3.7 Derived Measures

From our models, we derive estimates for three preregistered metrics that we use to compare visualization designs.

Linear in log odds (LLO) slopes measure the degree of bias in probability of superiority $\Pr(S)$ estimation (Fig. 3). A slope of one indicates unbiased performance, and slopes less than one indicate the degree to which users underestimate effect size.[5] We measure LLO slopes because they are very sensitive to the expected pattern of bias in responses, giving us greater statistical power than simpler measures like accuracy. Specifically, LLO slope is the expected increase in a user's logit-transformed probability of superiority estimate, $\text{logit}(response_{\Pr(S)})$, for one unit of increase in logit-transformed ground truth, $\text{logit}(true_{\Pr(S)})$. Using a linear metric (i.e., slope in logit-logit space) to describe an exponential response function in probability units comes from a theory that the brain represents probabilities on a log odds scale [18, 62]. The LLO model [18, 62] can be thought of as a generalization of the cyclical power model [23] that allows a varying intercept or a modification of Stevens' power law [49] for proportions.

[5] LLO slopes less than one represent bias toward the probability at the intercept, $\text{logit}^{-1}(intercept)$, which is close to $\Pr(S) = 0.5$ in our study.

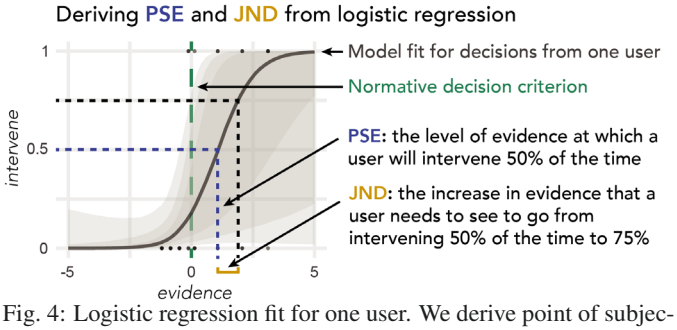


Fig. 4: Logistic regression fit for one user. We derive point of subjective equality (PSE) and just-noticeable difference (JND) by working backwards from probabilities of intervention to levels of evidence.

Points of subjective equality (PSEs) measure bias toward or against choosing to intervene in the decision task relative to a utility-optimal and risk-neutral decision rule (see Section 3.2). PSEs describe the level of evidence at which a user is expected to intervene 50% of the time (Fig. 4). A PSE of zero is utility-optimal, whereas a negative value indicates that a user intervenes when there is not enough evidence, and a positive value indicates that a user doesn't intervene until there is more than enough evidence. In our model, PSE is $\frac{-intercept}{slope}$ where $slope$ and $intercept$ come from the linear model in logistic regression.

Just noticeable-differences (JNDs) measure sensitivity to effect size information for the purpose of decision-making (Fig. 4). They describe how much additional evidence for the effectiveness of an intervention a user needs to see in order to increase their rate of intervening from 50% to about 75%. A JND in evidence units is a difference in the log probability of winning the award with the new player. We chose this scale for statistical inference because units of log stimulus intensity are thought to be approximately perceptually uniform [49, 56]. In our model, JND is $\frac{\text{logit}(0.75)}{slope}$ where $slope$ is the same as for PSE.

3.8 Participants

We recruited users through Amazon Mechanical Turk. Workers were located in the US and had a HIT acceptance rate of 97% or more. Based on the reliability of inferences from pilot data, we aimed to recruit 640 participants, 160 per uncertainty visualization. We calculated this target sample size by assuming that variance in posterior parameter estimates would shrink by a factor of roughly $\frac{1}{\sqrt{n}}$ if we collected a larger data set using the same interface. Since we based our target sample size on between-subjects effects (e.g., uncertainty visualization), our estimates of within-subjects effects (e.g., adding means) were very precise.

We recruited 879 participants. After our preregistered exclusion criterion that users needed to pass both attention checks, we slightly exceeded our target sample size with 643 total participants. However, we had issues fitting our model for an additional 21 participants, 17 of whom responded with only one or two levels of probability of superiority and 4 of whom had missing data. After these non-preregistered exclusions, our final sample size was 622 (with block order counterbalanced). All participants were paid regardless of exclusions, on average receiving $2.24 and taking 16 minutes to complete the experiment.

3.9 Qualitative Analysis of Strategies

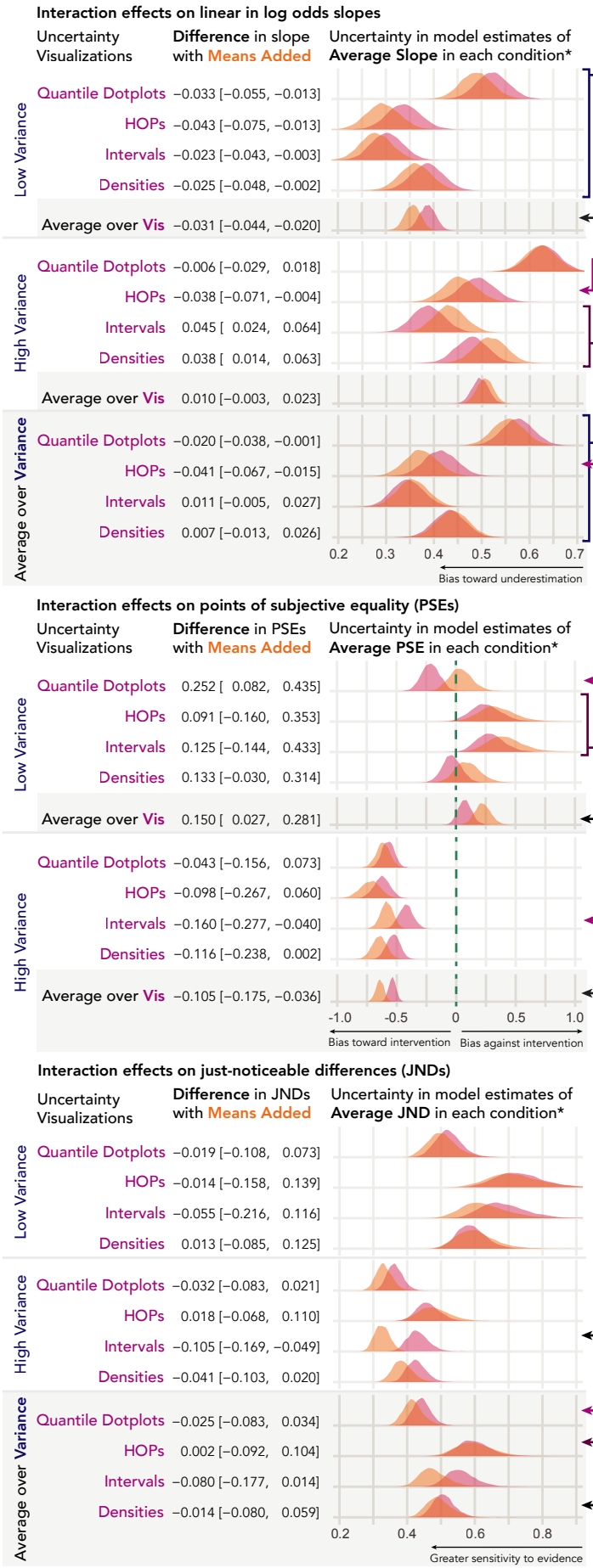
Using the two strategy responses we elicited from each user, we conducted a qualitative analysis to characterize users' visual reasoning strategies based on heuristics they used with different visualization designs (with and without means) and whether they switched strategies.

The first author developed a bottom-up open coding scheme for how users described their reasoning with the charts. Since some responses were uninformative about what visual properties of the chart a user considered (e.g., *"I used the charts to estimate the value added by the new player."*), we omitted participants for whom both responses were uninformative from further analysis. Excluding 180 such participants resulted in a final sample of 442 for our qualitative analysis.

We used our open codes to develop a classification scheme for strategies based on what visual features of charts users mentioned, whether they switched strategies, and whether they were confused by the chart or task. We coded for the following uses of visual features:

- **Relative position** of distributions
- **Means**, whether users relied on or ignored them
- **Spread** of distributions, whether users relied on variance, ignored it, or erroneously preferred high or low variance
- **Reference lines**, whether users relied on imagined or real vertical lines (e.g., the annotated decision threshold in Fig. 1 & 2)
- **Area**, whether users relied on the spatial extent of geometries
- **Frequencies**, whether users of quantile dotplots or HOPs relied on frequencies of dots or animated draws

Thus, we generated a spreadsheet of quotes, open codes, and categorical distinctions which enabled us to provide aggregate descriptions of patterns and heterogeneity in user strategies.



4 Results

4.1 Probability of Superiority Judgments

For each uncertainty visualization, **adding means** at **low variance** decreases LLO slopes. Recall that a slope of one corresponds to no bias, and a slope less than one indicates underestimation. When we **average over uncertainty visualizations**, **adding means** at **low variance** reduces LLO slopes for the average user, indicating a very small 0.8 percentage points increase in probability estimation error.

At **high variance**, the effect of **adding means** changes directions for different uncertainty visualizations. **Adding means** decreases LLO slopes for **HOPs**, whereas **adding means** increases LLO slopes for **intervals and densities**. Because differences in LLO slopes represent changes in the exponent of a power law relationship, these slope differences of similar magnitude indicate a very small increase in probability of superiority estimation error of 0.3 percentage points for HOPs and small reductions in error of about 1.5 and 1.0 percentage points for intervals and densities, respectively.

Users of all uncertainty visualizations underestimate effect size. When we **average over variance**, users show an average estimation error of 8.6, 14.0, 14.8, and 12.4 percentage points in probability of superiority units for quantile dotplots, HOPs, intervals, and densities, respectively, each **without means**. In this marginalization, **adding means** only has a reliable impact on LLO slopes for **HOPs**, but the difference is practically negligible.

4.2 Intervention Decisions

4.2.1 Points of Subjective Equality

For each uncertainty visualization, **adding means** at **low variance** increases PSEs. This results in different effects depending on whether the visualization with **no means** has a PSE below or above utility-optimal. Recall that a PSE of zero is utility-optimal, a negative PSE indicates intervening too often, and a positive PSE indicates not intervening often enough. Users of **quantile dotplots** with **no means** have negative PSEs which become unbiased when we **add means**. Users of **HOPs and intervals** with **no means** have positive PSEs, biases which increase when we **add means**. Users of **densities** with **no means** have PSEs near zero and become more biased when we **add means**. Only the effect for quantile dotplots is reliable. When we **average over uncertainty visualizations**, at **low variance** the average user may have a PSE 0.6 percentage points above utility-optimal with **no means**, and **adding means** increases this mild bias by about 1.7 percentage points in terms of the probability of winning.

At **high variance**, **adding means** decreases PSEs. Since PSEs for all uncertainty visualizations with **no means** are below optimal, **adding means** increases biases in all conditions, however, the effect is only reliable for **intervals**. When we **average over uncertainty visualizations**, at **high variance** the average user has a negative PSE 9.5 percentage points below utility-optimal with **no means**, and **adding means** increases this bias by about 2.1 percentage points.

4.2.2 Just-Noticeable Differences

At **low and high variance**, the effects of **adding means** on JNDs are mostly unreliable. Recall that smaller JNDs indicate that a user is sensitive to smaller differences in effect size for the purpose of decision-making. **Adding means** only has a reliable effect on JNDs for **intervals** at **high variance**, where it reduces JNDs by 1.2 percentage points in terms of the probability of winning.

When we **average over variance**, **quantile dotplots with means** lead to the smallest JNDs, and users of **HOPs with** or **without means** have the largest JNDs, a difference of about 1 percentage point in terms of the probability of winning. Quantile dotplots **with** or **without means** have reliably smaller JNDs than other conditions, with the exception of unreliable differences between quantile dotplots with **no means** and **densities with** or **without means**.

*Probability densities of model estimates show posterior distributions of means conditional on the average participant.

4.3 Discussion

Among the uncertainty visualizations we tested, quantile dotplots lead to the least biased probability of superiority estimates. This is not surprising given previous work (e.g., [17, 20, 26, 33, 34]) showing that frequency-based visualizations are effective at conveying probabilities. However, it is surprising that users do not perform reliably differently with frequency-based HOPs than with intervals or densities. HOPs directly encode probability of superiority by how often the draws from the two distributions change order, whereas in all other conditions users would need to calculate effect size analytically from visualized means and variances to arrive at the "correct" inference, although we doubt that users engage in such explicit mathematical reasoning. In Section 5, we present descriptive evidence of heuristics that users employ with different visualization designs, which helps to explain these results.

In most cases, the small effects on LLO slopes when adding means to uncertainty visualizations are probably negligible. However, they are consistent with the pattern of behavior we expect if users rely on visual distance between distributions as a proxy for effect size. When variance is lower relative the axis scale, distances between distributions look small even for large effects (Fig. 2, top), and users tend to underestimate effect size *more* when means are added. When variance is higher relative the axis scale, distances between distributions roughly correspond to effect size (Fig. 2, bottom), and users tend to underestimate effect size *less* when means are added, at least for densities and intervals.

Our results suggest that the best visualization design for utility-optimal decision-making probably depends on the level of variance relative to the axis scale. At lower variance, when multiple levels of variance are shown on a common scale, densities without means or quantile dotplots with means lead to the least bias in decisions. At higher variance, users are biased toward intervening in all conditions, and both densities without means and intervals without means lead to the least bias. The impact of means also depends on variance and axis scaling, such that when we average across uncertainty visualizations, adding means exacerbates biases that exist when means are absent. The effect of variance on PSEs (see Supplemental Materials) is large, such that users intervene more often at higher variance than at lower variance. One possible explanation for this is that users rely on distance between distributions as a proxy for effect size and make decisions as if effects are larger when distributions are further apart (Fig. 2).

Reported effects of visualization design on JNDs may not be practically important. All differences in JNDs between visualization designs are smaller than the difference between high versus low variance (see Supplemental Material). Smaller JNDs at high variance may reflect the fact that our high variance charts use white space more efficiently.

4.4 Comparing Magnitude Estimation & Decision-Making

Different visualization designs lead to the best performance on our magnitude estimation and decision-making tasks. To explore this decoupling of performance across tasks, we calculate average posterior estimates of our derived measures—LLO slope, PSE, and JND—for each individual user and compare them. Figure 5 shows that many individuals who are poor at magnitude estimation (i.e., LLO slopes below one) do well on the decision task (i.e., PSEs and JNDs near zero).

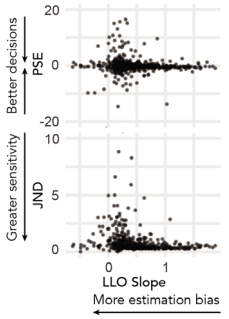


Fig. 5: PSEs and JNDs vs LLO slopes per user.

One possible explanation for this decoupling of performance on our two tasks is that users may rely on different heuristics to judge the same data for different purposes. This is consistent with Kahneman and Tversky's [31] distinction between *perceiving* the probability of an event to be p and *weighting* the probability of an event in decision-making as $\pi(p)$, which suggests that decision weights reflect preferences based on probabilities and risk attitudes [58]. Recent work in behavioral economics [35] suggests that biases in decision-making are partially attributable to imprecision in an individual's subjective perception of numbers (i.e., "number sense"). Since JNDs reflect the precision of perceived effect size implied by one's decisions and PSEs represent bias in decision-making, we can investigate this relationship within individual users in our study (Fig 6). In agreement with prior work, we see that greater sensitivity to effect size for decision-making (i.e., JNDs close to zero) predicts more utility-optimal decisions (i.e., PSEs close to zero). Although, based on the decoupling of LLO slopes and JNDs, it also seems clear that a user's internal sense of effect size is not necessarily identical when they use the same information for different tasks. We should be mindful that perceptual accuracy may not feed forward directly into decision-making.

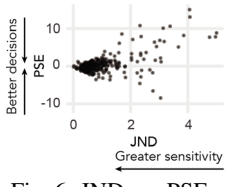


Fig. 6: JNDs vs PSEs.

5 Visual Reasoning Strategies

We use qualitative analysis of reported strategies to identify ways that users judge effect size by comparing distributions, giving us a vocabulary for how visualization design choices impact their interpretations.

5.1 Prevalent Strategies

The strategies we identify are not mutually exclusive. We count a user as employing a strategy if they mention it in either of their responses.

Only Distance: About 62% of users (275 of 442) rely on "how far to the right" the red distribution is compared to the blue one *without mentioning that they incorporate the variance of distributions into their judgments* (Fig. 2). Roughly 69% of these users (190 of 275) describe making a gist estimate of distance between distributions, with 46% (126 of 275) saying they rely on the mean difference specifically, and 13% (36 of 275) saying they rely on both gist distance and mean difference. Strategies which involve only the distance between distributions should result in a large bias toward underestimating effect size, which is what we see in our aggregated quantitative results.

Distance Relative to Variance: Only about 8% of users (35 of 442) mention that their interpretations of distance depend on the spread of distributions, suggesting that perhaps very few untrained users are sensitive to the impact of variance on effect size. If users estimate standard deviation and mean difference between distributions, they could use this information to calculate effect size analytically. However, we think it is far more likely that these users judge the distance between distributions relative to the spatial extent of uncertainty visualizations, which should result in underestimation bias which is similar to but less pronounced than with judgments of *only distance*.

Cumulative Probability: A substantial 36% of users (160 of 442) estimate the cumulative probability of winning the award with and/or without the new player. This strategy involves judging the distance, proportion of area, or frequency of markings across the threshold number of points to win (e.g., Fig. 7). These users may be confusing cumulative probability of winning the award, which is the best cue in the decision task, with probability of superiority (i.e., probability that team does better with the new player than without), which is what we ask for in the estimation task. However, since the probability of winning increases monotonically with probability of superiority, this strategy should theoretically result in milder underestimation bias than distance-based strategies.

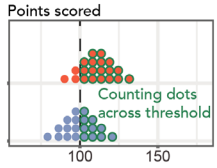


Fig. 7: Cumulative probability strategy with quantile dotplots.

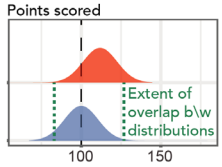


Fig. 8: Overlap strategy with densities.

Distribution Overlap: About 7% of users (31 of 442) describe judging the overlap between distributions. While similar to distance-based strategies, users conceptualize this strategy in terms of area rather than the gap between distributions (Fig. 8). For example, one user said they use HOPs *"only to see how much of an overlap [there is] between the two areas,"* suggesting that they imagine contours of distributions over the sets of animated draws. This strategy probably results in underestimation bias similar to judging *distance relative to variance*.

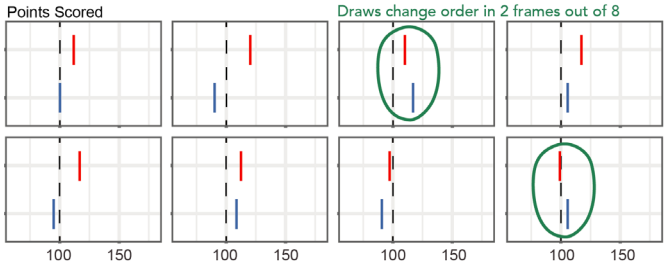


Fig. 9: Frequency of draws changing order strategy with HOPs.

Frequency of Draws Changing Order: This strategy is only relevant to the HOPs condition, where only about 16% of users (19 of 121) employed it. It involves judging the number of animated frames in which the draws from the two distributions switch order (Fig. 9). This is the best way to estimate probability of superiority from HOPs [26]. If we think of the user as accumulating information across frames, the precision of their inference is mostly limited by the number of frames they watch. For example, in Figure 9 red scores higher than blue in 6 of the 8 frames, and watching only 8 frames limits the precision of this inference to increments of $\frac{1}{8}$. The fact that only a handful of HOPs users employ this strategy helps to explain why the performance of HOPs users is worse than expected.

Switching Strategies: A substantial 29% of users (129 of 442) switch between strategies in the middle of the task. For example, one user of intervals without means described a mix of *cumulative probability* and *distribution overlap* strategies: *"If the red [distribution] was completely past the dotted line then I would buy the new player no matter what. If there were overlaps with blue I would just risk assess to see if it was worth it to me or not."* While more of a meta-strategy, our observation that a significant proportion of users switch is important because it suggests that judgment processes involved in graphical perception may not be consistent within each user.

5.2 Impacts of Visualization Design Choices

Users rely on visual features (Section 3.9) and strategies (Section 5.1) to varying degrees depending on visualization design (Table 1).

Intervals: Roughly 75% of intervals users (85 of 112) rely on *relative position* as a visual cue for effect size compared to 69% with densities (68 of 99), 61% with HOPs (74 of 121), and 59% with quantile dotplots (65 of 110). Of intervals users who look at *relative position*, about 87% (74 of 85) employ an *only distance* strategy, while only about 13% (11 of 85) judge *distance relative to variance* . In other words, only about 10% of intervals users (11 of 112) incorporate variance into their judgments of distance. About 28% of intervals users (31 of 112) report looking at *area*, with about 55% of these users (17 of 31) employing a *distribution overlap* strategy.

HOPs: About 61% of HOPs users (74 of 121) look at *relative position* to judge effect size. Of HOPs users who rely on *relative position*, merely 3% (2 of 74) use a *distance relative to variance* strategy. However, looking at *relative position* is not mutually exclusive with looking at *frequency* of draws, which 45% of HOPs users (54 of 121) rely on as a visual feature. Among HOPs users who rely on frequencies, about 69% (37 of 54) employ a *cumulative probability* strategy, while about 35% (19 of 54) rely on the optimal strategy of counting the *frequency of draws changing order*. Roughly 40% of HOPs users (48 of 121) mention *switching strategies* compared to 31% with intervals (35 of 112), 23% with densities (23 of 99), and 21% with quantile dotplots (23 of 110). Among HOPs users who switch strategies, about 81% (39 of 48) rely on the mean as a cue. Strategy switching involves the mean for about 30% of HOPs users who rely on *relative position* (22 of 74) compared to 43% of HOPs users who rely on *frequency* (23 of 54). That most HOPs users rely on *relative position*, and that those who do rely on *frequency* are more likely to switch to or from relying on the mean, helps to explain poor performance with HOPs.

Table 1: Frequency of strategies used per uncertainty visualization.

Strategy	Intervals	HOPs	Densities	Dotplots	Overall
Distance	73	77	61	64	275
Rel. to Var.	11	9	10	5	35
Cumulative	34	50	30	46	160
Overlap	17	2	9	3	31
Draw Order	0	19	0	0	19
Switching	35	48	23	23	129
Total	112	121	99	110	442

Densities: About 69% of densities users (68 of 99) rely on *relative position* as a visual cue. Of densities users who look at *relative position*, only about 13% (9 of 68) employ a *distance relative to variance* strategy. As one might expect, a substantial 36% of densities users (36 of 99) rely on *area* as a cue, compared to 10% of quantile dotplots users (11 of 110). Among densities users who rely on *area*, about 53% (19 of 36) employ a *cumulative probability* strategy, while about 28% (10 of 36) employ a *distribution overlap* strategy. Interestingly, about 27% of densities users (27 of 99) mention relying on the *spread* of distributions as a cue, more than the 21% of users with intervals (24 of 112), 21% with HOPs (25 of 121), and 10% with quantile dotplots (11 of 110) who report relying on the same cue.

Quantile Dotplots: Roughly 59% of quantile dotplots users (65 of 110) describe looking at *relative position* to judge effect size, similar to 61% of users with HOPs (74 of 121) and less than the 69% of densities users (68 of 99) and 76% of intervals users (85 of 112) who report using the same cue. Merely 6% of quantile dotplots users who rely on *relative position* (4 of 65) employ a *distance relative to variance* strategy. 37% of quantile dotplots users (41 of 110) rely on *frequency* as a visual cue by counting dots. About 81% of quantile dotplots users who rely on *frequency* (33 of 41) employ a *cumulative probability* strategy.

Adding Means: A substantial 35% of users (155 of 442) describe relying on the mean as a cue for effect size. If we split users based on whether or not they start the task with means, about 31% of users (67 of 218) *switch strategies* when means are added to the charts halfway through the task, compared to 10% (23 of 224) who switch strategies when means are removed. This asymmetry in strategy switching suggests that means are "sticky" as a cue: Among the 15% of users (67 of 442) who start with and rely on means, about 66% (44 of 67) attempt to visually estimate means *after means are removed from charts*, almost twice as many as the 34% (23 of 67) who switch to relying on other cues. However, the impact of adding means on performance depends on what other strategies a user is switching between. Among the 20% of users (90 of 442) who rely on means and switch strategies, about 44% (40 of 90) just incorporate the mean into judgments of *relative position* without relying on other visual cues. Other groups of users switch between relying on means and less similar visual cues, with 34% (31 of 90) also mentioning *frequency* and 12% (11 of 90) mentioning *area*. That many users switch between relying on *relative position* and means, and that strategies are heterogeneous, helps to explain why the average impact of means on performance is small in our results.

6 General Discussion

Our results suggest that *design guidelines for visualizing effect size* should depend on the user's task, the variance of distributions, and design choices about axis scales. To provide concrete design guidelines while acknowledging the inherent complexity of our results, we present high-level take-aways for designers alongside relevant caveats.

Quantile dotplots support the most perceptually accurate distributional comparisons, at least among the visualization designs we tested. **Caveat:** Asking users to perform two tasks may have led users to rely on relatively simple strategies like *cumulative probability* more than strategies which require more mental energy like *frequency of draws changing order*. Conditions of high cognitive load seem to favor uncertainty visualizations like quantile dotplots over HOPs.

Densities without means seem to support the best decision-making across levels of variance. On a fixed axis scale, densities without means and quantile dotplots with means perform best at lower variance, while densities without means and intervals without means perform best at higher variance. No visualization design we tested

eliminated bias in decision-making at higher variance. **Caveats:** The visualization design that leads to the least bias in decision-making depends on the variance of distributions relative to axis scale. Future work should investigate bias in decision-making over a gradient of variances shown on a common scale, including charts with heterogeneous variances, as this would enable more exhaustive design recommendations.

Adding means leads to small biases in magnitude estimation and decision-making from distributional comparisons, leading users to underestimate effect size and make less utility-optimal decisions in most in most cases we tested. **Caveats:** Although the biasing effects of means are mostly negligible, our estimates of these biases are probably very conservative for two reasons: (1) added means were only highly salient in the HOPs condition; and (2) in the absence of added means, users already tend to rely on *relative position*, a cue which the mean merely reinforces. The effects of adding means on decision quality reverse at high versus low variance, so these biases may disappear for specific combinations of variance and axis scale.

Users rely on distance between distributions as a proxy for effect size, so designers should note when this will be misleading and encourage more optimal strategies. Our quantitative analysis shows that adding means induces small but reliable biases in magnitude estimation, consistent with distance-based heuristics. Our qualitative analysis of strategies verifies that the majority of users (357 of 442; 80.8%) rely on distance between distributions or mean difference to judge effect size. **Caveats:** Subtle design choices probably impact the tendency to rely on distance heuristics versus other strategies. For example, including a decision threshold annotation on our charts (Fig. 2) may have encouraged users to judge effect size as *cumulative probability*, rather than probability of superiority, contributing to underestimation bias.

6.1 Limitations

We only tested symmetrical distributions, and this may limit the generalizability of our inferences. Although we speculate that chart users may rely on central tendency regardless of the family of a distribution, reasoning with multi-modal distributions in particular may involve different strategies not accounted for in the present study.

Because we rely on self-reported strategies in our qualitative analysis, our findings only reflect *conscious* strategies. This leaves out implicit or automatic information processing such as visual adaptation [32] and ensemble processing [51], except in rare cases where users report trying to *"roughly average"* predictions presented as HOPs.

Our choice to incentivize the decision-making task but not magnitude estimation may have contributed to the decoupling of performance on our two tasks. We cannot disentangle this possible explanation from evidence corroborating Kahneman and Tversky's [31] distinction between perceived probabilities and decision weights (see Section 4.4).

We control the incentives for our decision task rather than manipulating them, in part because it is not feasible to test dramatically different incentives on Mechanical Turk. As such the risk preferences that we measure as PSEs are representative of users optimizing small monetary bonuses, and they may not capture how people respond to visualized data in crisis situations when lives, careers, or millions of dollars are at stake. However, by devising a task that is representative of a broad class of decision problems (see Section 3.2), we make our results as broadly applicable as possible. We speculate that the *relative impacts* of visualization designs on risk preferences should generalize to decision problems with similar utility functions.

6.2 Satisficing and Heterogeneity

The visual reasoning strategies that chart users rely on when making judgments from uncertainty visualizations may not be what visualization designers expect. We present evidence that, in the absence of training, users satisfice by using suboptimal heuristics to decode the signal from a chart. We also find that not all users rely on the same strategies and that many users switch between strategies. Satisficing and heterogeneity in heuristics make it difficult both to anticipate how people will read charts and to study the impact of design choices. Conventionally, visualization research has characterized visualization effectiveness by ranking visualization designs based on the performance of the average user (e.g., [7]). However, in cases like the present study where users are heterogeneous in their strategies, these *averages may not account for the experience of very many users* and are probably an oversimplification. Visualization researchers should be mindful of satisficing and heterogeneity in users' visual reasoning strategies, attempt to model these strategies, and try to design ways of training users to employ more optimal strategies.

6.3 Toward Better Models of Visualization Effectiveness

Because some users seem to adopt suboptimal strategies or switch between strategies when presented with an uncertainty visualization, models of visualization effectiveness which codify design knowledge and drive automated visualization recommendation and authoring systems should represent these strategies. We envision a new class of behavioral models for visualization research which attempt to enumerate possible strategies, such as those we identify in our qualitative analysis, and learn how often users employ them to perform a specific task when presented with a particular visualization design. Previous work [27] demonstrates a related approach by calculating expected responses based on a set of alternative perceptual proxies for visual comparison and comparing these expectations to users' actual responses. Like the present study, this work describes the correspondence between expected patterns and user behavior. Instead, we propose incorporating functions representing predefined strategies into predictive models which estimate the proportion of users employing a given strategy.

In a pilot study, we attempted to build such a model: a Bayesian mixture model of alternative strategy functions. However, because multiple strategies predict similar patterns of responses, we were not able to fit the model due to problems with identifiability. This suggests that the kind of model we propose will only be feasible if we design experiments such that alternative strategies predict sufficiently different patterns of responses. The approach of looking at the agreement between proxies and human behavior [27] suffers the same limitation, but there is no analogous mechanism to identifiability in Bayesian models to act as a fail-safe against unwarranted inferences. Future work should continue pursuing this kind of strategy-aware behavioral modeling.

We want to emphasize that the proposed modeling approach is not strictly quantitative, as the definition of strategy functions requires a descriptive understanding of users' visual reasoning. As such this approach offers a way to formalize the insights of qualitative analysis and represent the gamut of possible user behaviors inside of visualization recommendation and authoring systems.

7 Conclusion

We contribute findings from a mixed design experiment on Mechanical Turk investigating how visualization design impacts judgments and decisions from effect size. Our results suggest that visualization designs which support the least biased estimation of effect size do not necessarily support the best decision-making. We discuss how a user's sense of the signal in a chart may not necessarily be identical when they use the same information for different tasks. We also find that adding means to uncertainty visualizations induces small but reliable biases consistent with users relying on visual distance between distributions as a proxy for effect size. In a qualitative analysis of users' visual reasoning strategies, we find that many users switch strategies and do not employ an optimal strategy when one exists. We discuss ways that canonical characterizations of graphical perception in terms of average performance gloss over possible heterogeneity in user behavior, and we propose opportunities to build strategy-aware models of visualization effectiveness which could be used to formalize design knowledge in visualization recommendation and authoring systems beyond context-agnostic rankings of chart types.

Acknowledgments

We thank the members of the UW IDL and Vis-Cog Lab, as well as the MU Collective at Northwestern for their feedback. This work was supported by a grant from the Department of the Navy (N17A-T004).

REFERENCES

[1] J. Baron. *Thinking and deciding (4th ed.)*. Cambridge University Press, 2008.

[2] N. J. Barrowman and R. A. Myers. Raindrop plots: A new way to display collections of likelihoods and distributions. *American Statistician*, 57(4):268–274, 2003. doi: 10.1198/0003130032369

[3] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods*, 10(4):389–396, 2005. doi: 10.1037/1082-989X.10.4.389

[4] W. C. Brinton. *Graphic Presentation*. Brinton Associates, 1939.

[5] P.-C. Bürkner. brms: Bayesian Regression Models using 'Stan', 2020.

[6] B. Chance, J. Garfield, and R. DelMas. Developing Simulation Activities to Improve Students ' Statistical Reasoning. In *Proceedings of the International Conference on Technology in Mathematics Education*, pp. 2—-10, 2000.

[7] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[8] R. Coe. It's the effect size, stupid: What effect size is and why it is important. 2002.

[9] M. Correll, E. Bertini, and S. Franconeri. Truncating the Y-Axis: Threat or Menace? In *ACM Human Factors in Computing Systems (CHI)*, 2020.

[10] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014. doi: 10.1109/TVCG.2014.2346298

[11] G. Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966

[12] G. Cumming and S. Finch. Inference by Eye. *American Psychologist*, 60(2):170–180, 2005. doi: 10.1037/0003-066X.60.2.170

[13] P. Cummings. Arguments for and Against Standardized Mean Differences (Effect Sizes). *Archives of Pediatrics and Adolescent Medicine*, 165(7):592–596, 2011.

[14] M. Fernandes, L. Walls, S. A. Munson, J. Hullman, and M. Kay. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *ACM Transactions on Computer-Human Interaction*, number April. Montreal, 2018. doi: 10.1145/3173574.3173718

[15] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 182(2):389–402, 2019. doi: 10.1111/rssa.12378

[16] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer. Using Icon Arrays to Communicate Medical Risks: Overcoming Low Numeracy. *Health Psychology*, 28(2):210–216, 2009. doi: 10.1037/a0014474

[17] G. Gigerenzer and U. Hoffrage. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological review*, 102(4):684–704, 1995. doi: 10.1037/0033-295X.102.4.684

[18] R. Gonzalez and G. Wu. On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38:129–166, 1999. doi: 10.1007/978-3-319-89824-7_85

[19] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5):1157–1164, 2014. doi: 10.3758/s13423-013-0572-3

[20] U. Hoffrage and G. Gigerenzer. Using natural frequencies to improve diagnostic inferences. *Academic medicine: Journal of the Association of American Medical Colleges*, 73(5):538–540, 1998. doi: 10.1097/00001888-199805000-00024

[21] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020. doi: 10.1145/3313831.3376454

[22] R. M. Hogarth and E. Soyer. Sequentially Simulated Outcomes: Kind Experience Versus Nontransparent Description. *Journal of Experimental Psychology: General*, 140(3):434–463, 2011. doi: 10.1037/a0023265

[23] J. G. Hollands and B. P. Dyre. Bias in Proportion Judgments: The Cyclical Power Model. *Psychological Review*, 107(3):500–524, 2000. doi: 10.1037//0033-295X.107.3.500

[24] D. Huff. *How to Lie with Statistics*. WW Norton & Company, 1993.

[25] J. Hullman. Why Authors Don't Visualize Uncertainty. In *IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers*, 2020.

[26] J. Hullman, P. Resnick, and E. Adar. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PloS one*, 10(11):e0142444, 2015. doi: 10.1371/journal.pone.0142444

[27] N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri. The Perceptual Proxies of Visual Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1012–1021, 2020. doi: 10.1109/TVCG.2019.2934786

[28] S. Joslyn and J. LeClerc. Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science*, 22(4):308–315, 2013. doi: 10.1177/0963721413481473

[29] S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(1):126–140, 2012. doi: 10.1037/a0025185

[30] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.

[31] D. Kahneman, A. Tversky, B. Y. D. Kahneman, and A. Tversky. Prospect Theory : An Analysis of Decision under Risk. *Econometrica*, 47(2):263–292, 1979.

[32] A. Kale and J. Hullman. Adaptation and learning priors in visual inference. In *VisXVision Workshop at IEEE VIS 2019*. Vancouver, BC, Canada, Oct 2019.

[33] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2019.

[34] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 ACM annual conference on Human Factors in Computing Systems*, 2016.

[35] M. W. Khaw, Z. Li, and M. Woodford. Risk Aversion as a Perceptual Bias. 2017.

[36] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman. A Bayesian Cognition Approach to Improve Data Visualization. In *ACM Human Factors in Computing Systems (CHI)*, 2019.

[37] C. F. Manski. Communicating uncertainty in policy analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 2018:201722389, 2018. doi: 10.1073/pnas.1722389115

[38] C. F. Manski. The Lure of Incredible Certitude. *Working Paper*, 2018.

[39] C. F. Manski. The lure of incredible certitude. *Economics & Philosophy*, pp. 1–30, 2019.

[40] K. O. McGraw and S. Wong. A common language effect size statistic. *Psychological bulletin*, 111(2), 1992.

[41] L. Micallef, P. Dragicevic, and J. Fekete. Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing to cite this version. *IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers*, 18(12):2536–2545, 2012.

[42] K. Pearson. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186:343—414, 1895.

[43] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, E. authors, S. Heisterkamp, B. Van Willigen, and R-core. nlme: Linear and Nonlinear Mixed Effects Models, 2020.

[44] T. Schäfer and M. A. Schwarz. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10(APR):1–13, 2019. doi: 10.3389/fpsyg.2019.00813

[45] H. A. Simon. Rational Choice and the Structure of the Environment. *Psychological Review*, 63(2):129–138, 1956. doi: 10.1037/xge0000013

[46] I. M. Sobol. Uniformly distributed sequences with an additional uniform property. *U.S.S.R. Comput. Maths. Math. Phys.*, 16:236–242, 1976.

[47] E. Soyer and R. M. Hogarth. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3):712–714, 2012. doi: 10.1016/j.ijforecast.2012.02.004

[48] D. J. Spiegelhalter. Surgical Audit: Statistical Lessons from Nightingale and Codman. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(1):45–58, 1999.

[49] S. S. Stevens. On the psychophysical law. *Psychological Review*, 64:153–181, 1957.

[50] E. Stoycheff. Please participate in Part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations*, 9:1–5, 2016. doi: 10.1177/2059799116672879

[51] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of

ensemble coding in data visualizations. *Journal of Vision*, 16(5), 2016.
[52] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical report, 1994.
[53] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley Pub, Reading, Mass, 1977.
[54] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.
[55] A. Tversky and D. Kahneman. The Framing Of Decisions And The Psychology Of Choice. *Science*, 211(30):453–458, 1981.
[56] L. R. Varshney and J. Z. Sun. Why do we perceive logarithmically? *Significance*, February:28–31, 2013.
[57] J. von Neumann, O. Morgenstern, and A. Rubinstein. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944.
[58] E. U. Weber. From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events. *Psychological Bulletin*, 115(2):228–242, 1994.
[59] S. Westwood, S. Messing, and Y. Lelkes. Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *Journal of Politics*, pp. 1–38, 2019. doi: 10.2139/ssrn.3117054
[60] G. N. Wilkinson and C. E. Rogers. Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):392–399, 1973.
[61] J. K. Witt. Introducing hat graphs. *Cognitive Research: Principles and Implications*, 4(1), 2019. doi: 10.1186/s41235-019-0182-3
[62] H. Zhang, L. T. Maloney, and D. Lange. Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6(January):1–14, 2012. doi: 10.3389/fnins.2012.00001